Information Processing and Management 63 (2026) 104672

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Structure-aware multi-view urban representation learning with coordinated fusion and alignment

Jinghui Wei [a,b], Sheng Wu [b,c], Shifen Cheng [d,e,*], Peixiao Wang [d,e], Feng Lu [b,d,e]

[a] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[b] The Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350108, China
[c] Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou 350108, China
[d] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[e] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Urban representation learning
Multi-view data fusion
Contrastive learning
Embedding
Coordinated optimization

ABSTRACT

Urban representation learning leverages heterogeneous data. While unified frameworks that combine feature fusion and contrastive learning have achieved promising results, two key challenges remain: 1) the lack of structural awareness often leads to suboptimal negative sampling and reduces the discriminability of embeddings; and 2) the inherently conflicting optimization objectives between fusion and contrastive modules may result in unstable training and suboptimal convergence. To address these issues, we propose SAMC, a **S**tructure-**A**ware **M**ulti-view representation learning framework with **C**oordinated fusion and alignment. SAMC introduces a multi-view contrastive learning module that incorporates structural similarity into negative sampling, thereby enhancing semantic coherence and cross-view consistency. To improve training stability, a flexible optimization strategy that incorporates soft Lagrangian constraints and stepwise state tracking is designed to coordinate gradient updates across fusion and alignment modules. Experiments on multiple urban datasets show that SAMC achieves average improvements of approximately 17.2%, 14.2%, and 5.9% compared to the state-of-the-art baselines on the tasks of regional popularity prediction, service demand prediction, and land use classification, respectively. Visualization analyses further confirm that SAMC enhances the discriminability, robustness, and generalizability of urban representations by embedding structural priors and adopting multi-stage coordinated optimization. Moreover, SAMC achieves a favorable trade-off between computational efficiency and predictive performance.

1. Introduction

With the increasing availability of sensing technologies, representation learning has emerged as a key technique in urban computing. It aims to map urban spatial regions into a low-dimensional space while preserving their functional and semantic

* Corresponding author at: State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.
E-mail addresses: 231016005@fzu.edu.cn (J. Wei), wusheng@fzu.edu.cn (S. Wu), chengsf@lreis.ac.cn (S. Cheng), wpx@lreis.ac.cn (P. Wang), luf@lreis.ac.cn (F. Lu).

https://doi.org/10.1016/j.ipm.2026.104672
Received 29 July 2025; Received in revised form 1 February 2026; Accepted 1 February 2026
Available online 6 February 2026

characteristics (Deng et al., 2024). Such technique has been widely adopted in tasks such as population density estimation (Neal et al., 2022; Zhang et al., 2024), land use classification (Cao et al., 2025a; Wu et al., 2022; Xu et al., 2023), and crime pattern analysis (Fu et al., 2025; Vomfell et al., 2018; Zhou et al., 2023). These applications support the modeling and reasoning of complex urban systems (Chen et al., 2025; Zou et al., 2025).

Existing embedding methods for urban applications are generally divided into single-view and multi-view approaches. Single-view methods typically rely on a single type of urban data, such as human trajectories (Yao et al., 2018), points of interest (POI) data (Luo et al., 2022; Wang et al., 2018), or remotely sensed images (Jean et al., 2019), to construct region-level representations. These methods are computationally efficient but overlook the complementary nature of diverse data sources, which limits their ability to fully represent regional semantics. In contrast, multi-view embedding methods integrate heterogeneous data sources to learn regional features from different perspectives. They improve representation quality by leveraging complementary strengths across modalities, which in turn enhances performance on downstream tasks (Chan et al., 2023; Deng et al., 2024; Li et al., 2024a).

Current multi-view embedding approaches generally follow two paradigms: feature fusion and representation alignment. Feature fusion first encodes representations from each view using models such as graph attention networks (Wang et al., 2022a) or autoencoders (Li et al., 2024a; Luo et al., 2022). These representations are then integrated through strategies such as concatenation, weighted averaging, or graph-based fusion (Cao et al., 2025b; Sun et al., 2024; Yao et al., 2017). This integration aims to enhance representational expressiveness by exploiting cross-view complementarity (Chan et al., 2023; Wang et al., 2020). Representation alignment, in contrast, focuses on modeling consistency across views. It commonly leverages contrastive learning to construct positive and negative sample pairs across views. Specifically, positive pairs are formed using representations of the same region from different views, while negative pairs are formed using representations from different regions. This design encourages representations of the same region to align in a shared latent space while preserving inter-region discriminability (Li et al., 2024b, 2023; Trosten et al., 2021; Xu et al., 2022; Zhang et al., 2022a). To combine the strengths of both paradigms, recent studies have proposed unified integration frameworks aiming to simultaneously capture complementary information across views, and enforce semantic consistency within a shared embedding space. Such a strategy demonstrates strong potential in improving representation quality and generalization performance (Wang et al., 2024; Yan et al., 2025; Zhang et al., 2022a). Despite these advances, two critical limitations remain underexplored.

Limitation 1: Lack of structural awareness in contrastive learning. In current unified integration frameworks, contrastive learning typically constructs cross-view positive and negative pairs, where positive samples represent the same region from different views, and negative samples are randomly selected from other regions. However, in real urban environments, regions with different spatial locations may still share similar functional structures or semantic roles. For example, business districts with comparable POI compositions, mobility flows, or socioeconomic patterns may play analogous roles in the urban systems. Treating such semantically similar regions as negative pairs introduces misleading supervision, thereby hindering the model's ability to recognize shared functional characteristics across geographically distant regions. Consequently, contrastive learning that ignores structural similarity often fails to capture latent semantic connections, reducing the compactness and discriminative quality of the representations.

Limitation 2: Optimization conflicts between fusion and alignment. Feature fusion typically seeks to minimize differences in shared semantic representations across views to enhance semantic consistency. In contrast, contrastive learning aims to maximize the semantic separation between urban representations to improve discrimination and generalization. These two objectives inherently conflict in optimization directions. Without proper coordination, fusion modules tend to pull all views closer, whereas contrastive modules push apart representations of structurally dissimilar regions. This leads to conflicting gradient updates, which cause unstable convergence and suboptimal embedding structures. Therefore, effectively reconciling these conflicting gradients remains a central challenge in unified fusion-alignment frameworks for achieving robust and generalizable region embeddings.

To address these limitations, we propose a novel framework named SAMC (***S**tructure-**A**ware **M**ulti-view representation learning with **C**oordinated fusion and alignment*), which jointly enhances structural awareness and resolves the optimization conflicts between fusion and contrastive objectives in multi-view representation learning. To overcome the lack of structural awareness, SAMC introduces a structure-guided contrastive module that integrates functional similarity into the negative sampling. This design prevents semantically similar regions from being treated as negatives, thereby improving semantic consistency and cross-view alignment. To mitigate the optimization conflict between fusion and contrastive modules, SAMC adopts a coordinated training strategy to harmonize conflicting gradients, ensuring more stable convergence and higher-quality embeddings. In general, the main contributions of this study are as follows:

(1) **Structure-aware contrastive learning:** We propose a contrastive learning mechanism that explicitly incorporates latent structural relations among regions into the negative sampling process. By avoiding semantically misleading supervision, it improves the cohesion of structurally similar regions and enhances the discriminative power of the embedding process.
(2) **Multi-stage coordinated optimization with soft constraints:** We develop a novel optimization strategy that combines modular training with soft Lagrangian regularization. This approach effectively coordinates fusion and contrastive learning objectives, alleviates gradient conflicts, and enhances the stability and efficiency of joint optimization.
(3) **Comprehensive empirical validation:** Extensive experiments conducted on multiple real-world datasets demonstrate that SAMC consistently outperforms state-of-the-art baselines across various downstream tasks. These results validate the effectiveness of the proposed framework in integrating heterogeneous data while maintaining structural and semantic consistency in urban embeddings.

2. Related work

2.1. Urban representation learning

Urban representation learning aims to generate low-dimensional vector embeddings of urban regions that encapsulate their structural and functional characteristics (Cao et al., 2025b). These embeddings provide a compact yet informative representation, supporting a wide range of downstream applications, such as land-use classification, population estimation, and environmental forecasting (Yong et al., 2024; Zhang et al., 2019; Zhao et al., 2025). Existing approaches can be broadly categorized into single-view and multi-view methods based on the nature of the input data.

Single-view methods focus on learning region embeddings from a single data source. Common sources include human mobility data (Wang et al., 2017; Yao et al., 2018; Zhang et al., 2020) and regional attribute data (Jean et al., 2019; Jin et al., 2024; Li et al., 2023). These methods typically fall into three main paradigms: (1) graph-structured approaches, such as Hierarchical Graph Infomax (HGI) (Huang et al., 2023) and HyperRegion (Deng et al., 2024); (2) feature aggregation methods, including Tile2Vec (Jean et al., 2019) and Dual Contrastive Learning (RegionDCL) (Li et al., 2023); and (3) spatiotemporal feature modeling methods, such as Mobility-based Zone Embedding (ZE-Mob) (Yao et al., 2018) and Multi-Graph Fusion Networks (MGFN) (Wu et al., 2022). While effective in certain scenarios, single-view methods are inherently constrained by the limitations of a single modality, which reduces their capacity to comprehensively characterize regional complexity and limits their generalizability across diverse urban contexts.

To overcome these limitations, multi-view approaches have been increasingly adopted. They integrate heterogeneous data sources to produce more robust and comprehensive embeddings (Chan et al., 2023, 2024b; Fang et al., 2021; Li et al., 2018; Robinson et al., 2017; Wang et al., 2024). These methods can be broadly grouped into two main paradigms: feature fusion and representation alignment (Wu et al., 2019; Xiao et al., 2021). Among them, feature fusion-based methods combine complementary information from multiple views into a unified embedding by explicitly modeling inter-view interactions. For example, Multi-View representation learning for Urban Region Embedding (MVURE) (Zhang et al., 2020) constructs a joint multi-view graph with a shared graph structure constraint to facilitate cross-view feature integration. Similarly, MGFN (Wu et al., 2022) employs a multi-graph fusion network to capture multi-level interactions across views, which improves embedding stability. In contrast, representation alignment methods aim to preserve view-specific information while encouraging semantic consistency across views. A representative example is multi-view Contrastive Prediction model for urban Region embedding (ReCP) (Li et al., 2024b), which jointly optimizes view-specific and consensus representations through contrastive and reconstruction objectives, thereby achieving high inter-view consistency.

2.2. Contrastive learning methods

Contrastive learning aims to structure the latent space by minimizing the distance between semantically similar instances while maximizing the distance between dissimilar ones (Ma et al., 2024; Wang et al., 2025). In single-view settings, where multi-view data are unavailable, positive pairs are typically generated through data augmentation techniques applied to the same instance, whereas negative pairs consist of samples from different instances (Li et al., 2024b; Zhang et al., 2025). In multi-view settings, where distinct views offer complementary characterizations of the same object, embeddings from different views of the same object are treated as positive pairs, while those from different objects constitute negative pairs.

Recent studies have demonstrated the effectiveness of multi-view contrastive learning in tasks such as multi-view clustering and urban representation (Li et al., 2022; Wang et al., 2022b). For instance, Zhang et al. (2022a) proposed a joint intra-view and inter-view contrastive framework, in which intra-view contrastive learning enhances feature discrimination within each view, and inter-view contrastive learning enforces cross-view consistency to facilitate knowledge transfer. From an information-theoretic perspective, Li et al. (2024b) sought to reduce view-specific disparities by maximizing mutual information and minimizing conditional entropy across views, thereby improving the robustness and generalizability of the learned representations. The success of contrastive learning has been further extended to graph-structured domains, giving rise to graph contrastive learning (GCL) (Han et al., 2022). GCL leverages the intrinsic graph topology to construct positive and negative pairs, allowing it to capture multi-hop and higher-order relational semantics (Yu and Jia, 2024).

Despite these advances, existing approaches still face critical limitations. First, GCL's reliance on a predefined graph structure can constrain the model, hindering the discovery of latent relationships beyond the given topology. Second, in cross-view contrastive learning, the failure to account for structural similarity may cause semantically similar regions to be incorrectly treated as negative pairs. Such mislabeling introduces conflicting training signals, which hampers the model's ability to recognize shared functional patterns across spatially dispersed yet functionally similar regions. Consequently, the resulting embeddings may suffer from reduced compactness and discriminative power, thereby limiting their effectiveness in downstream tasks.

2.3. Joint learning methods

Recent advances have explored the integration of feature fusion and representation alignment to harness their complementary strengths: feature fusion aggregates diverse information across views, while contrastive alignment enforces semantic consistency. For example, CREME (Zhang et al., 2022b) employed a multi-view aggregator to integrate diverse representations and introduced two collaborative contrastive objectives—view fusion InfoMax and inter-view InfoMin—to support self-supervised training and enhance discriminative capability. Ke et al. (2023) proposed a clustering-guided fusion network to learn shared representations and designed an asymmetric contrastive strategy to prevent alignment between view-specific features, thus mitigating suboptimal solutions. Yu et al.

(2024) incorporated an inter-view contrastive module to capture cross-view commonalities and introduced an attention mechanism to assess the relative semantic contribution of each view, thereby facilitating more effective modeling of inter-view variation. Despite these advances, joint optimization remains challenging. The inherent conflict between fusion and alignment objectives may lead to an unstable training process and compromised representation quality.

3. Preliminary

In this section, we present the basic definitions, problem statement and objectives. For clarity and ease of reference, the commonly used symbols in this paper are summarized in Table 1.

3.1. Definitions and problem statement

Definition 1. **(Region):** Given a specific partitioning method—such as census tracts—a city is divided into a set of non-overlapping regions, denoted as $R = \{r_1, r_2, \ldots, r_n\}$, where r_i represents the i th region, and n denotes the total number of regions.

Definition 2. **(Regional semantic features):** Regional semantic features describe the functional, social, and categorical attributes of urban areas. These features are typically derived from Point of Interest (POI) data (Zhang et al., 2022a). Formally, let $P = \{p_1, p_2, \ldots, p_n\}$, where $p_i \in R^C$ denotes the semantic feature vector of region r_i. Each element of p_i represents the count of POIs in a specific category within region r_i, and C denotes the total number of POI categories.

Definition 3. **(Regional interaction features):** Regional interaction features characterize the spatial patterns of mobility and the behavioral dynamics of resource flows—such as population, capital, and information—across urban regions. In this study, we capture these features using inter-regional taxi trip data, which serve as proxies for the flow of resources between regions (Deng et al., 2024). To better reflect directionality, regional interaction features are decomposed into outflow and inflow components. Formally, let $S = \{s_1, s_2, \ldots, s_n\}$, where $s_i \in R^n$ denotes the outflow vector for region r_i. Each element of s_i represents the number of trips originating from region r_i to other regions. Similarly, let $D = \{d_1, d_2, \ldots, d_n\}$, where $d_i \in R^n$ denotes the inflow vector for region r_i. Each element of d_i represents the number of trips arriving at region r_i from other regions.

Problem statement: In this work, we aim to learn urban representations from three views ($V = 3$). Given the regional semantic features view P, as well as the regional interaction feature view S and D, the objective is to learn a low-dimensional embedding set $\varepsilon = \{E_1, E_2, \ldots, E_n\}$, where $E_i \in R^d$ represents the latent feature vector of region r_i, and d denotes the embedding size. These embeddings are designed to capture the intrinsic semantic and interactional characteristics of the regions in a shared latent space.

3.2. Research objectives

The research objectives (RO) of this paper are as follows:

(RO1) To design a novel contrastive learning mechanism that alleviates the loss of discriminative power caused by semantically misleading supervision.
(RO2) To develop a new training strategy that resolves the gradient conflict between fusion and contrastive learning objectives during optimization.
(RO3) To ensure that the proposed framework achieves high accuracy on downstream tasks while maintaining low computational cost.

Table 1
The commonly used symbols in this paper.

Symbol	Explanation
n	The total number of regions in the city.
P	The semantic feature matrix of regions in the city.
S	The interaction outflow feature matrix of regions in the city.
D	The interaction inflow feature matrix of regions in the city.
ε	The low-dimensional embedding set of regions in the city.
X^v	The original input data for the v-th view
$E_v(\cdot)$	The encoder of the v-th view.
$D_v(\cdot)$	The decoder of the v-th view.
H^v	The view-specific representation of the v-th view.
Z^v	The enhanced view-specific representation of the v-th view.
$\widehat{Z}_c$	The consensus representation output by the CVTF module.
$\widehat{Z}_m$	The enhanced consensus representation output by the MVSCL module.
L_r^v	The reconstruction loss of the v-th view.
L_{con}	The consistency loss.
L_c	The inter-view contrastive loss.

4. METHODOLOGY

4.1. Overall framework

The overall architecture of SAMC, as illustrated in Fig. 1, consists of four functional modules and a multi-stage optimization strategy. To facilitate the learning of high-quality consensus representations, SAMC first introduces three key modules: View Reconstruction (VR), Enhanced View-Specific Representation (EVSR), and Cross-View Transformer (CVTF). These modules collaboratively produce a set of denoised and complementary view-specific representations $Z = \{Z^1, Z^2, \ldots Z^v\}$, while effectively integrating multi-view information into a unified consensus representation $\widehat{Z}_c$. These representations provide a robust foundation for subsequent contrastive learning process.

Building upon these enriched representations, a Multi-View Structure-Aware Contrastive Learning (MVSCL) module is designed to refine consensus embedding. MVSCL employs a structure-guided negative sampling strategy, which leverages functional similarity to prevent semantically similar regions from being treated as negatives. This strategy not only preserves semantic coherence but also

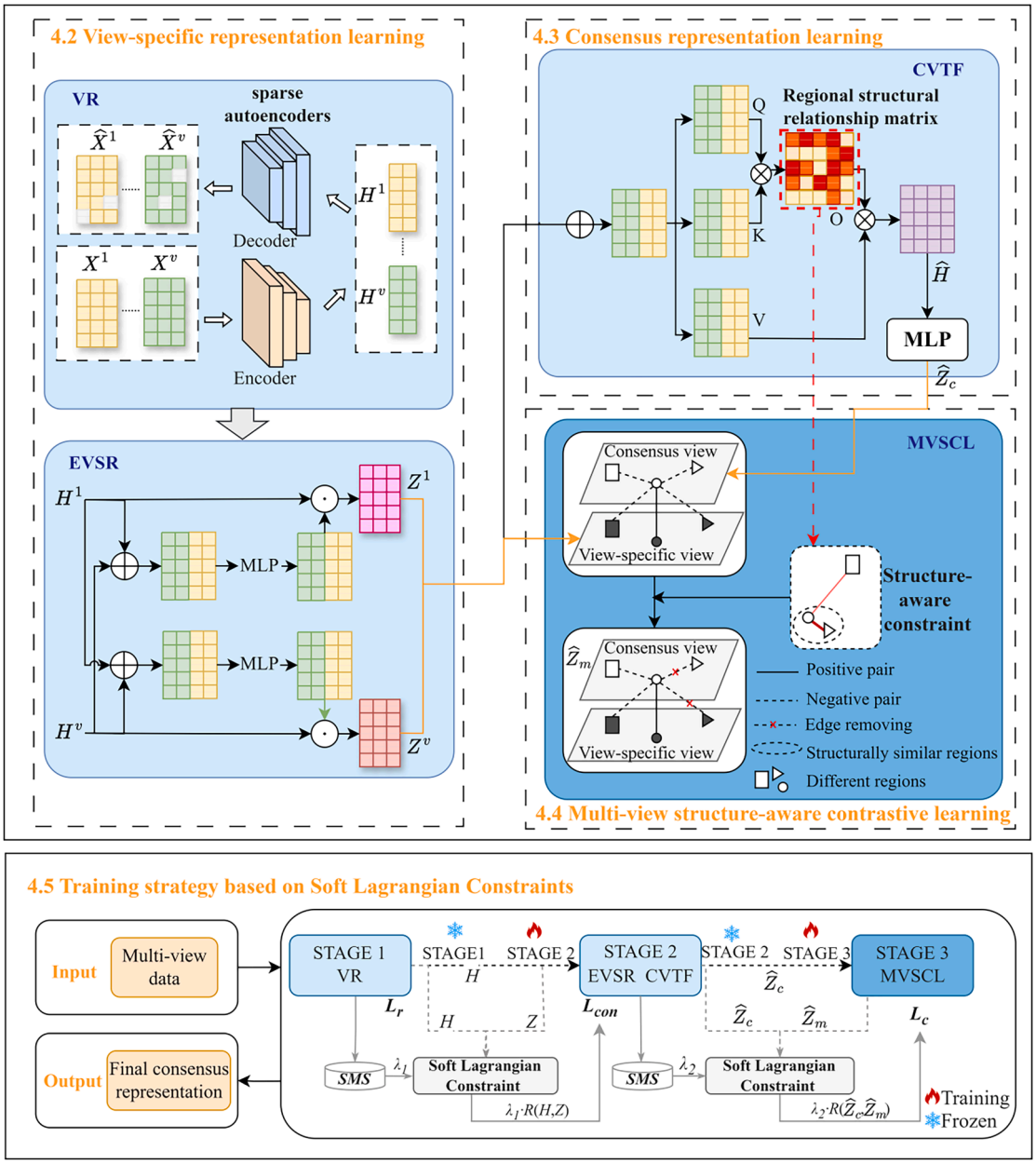


Fig. 1. Overall architecture of the proposed SAM model.

enhances the alignment between the consensus and view-specific embeddings, thereby improving the discriminability of functionally distinct regions.

To further enhance convergence stability and embedding quality, SAMC adopts a coordinated training strategy inspired by the Alternating Direction Method of Multipliers (ADMM) (Liu et al., 2024). This strategy decomposes the overall training process into modular sub-tasks and incorporates soft Lagrangian constraints to maintain consistency across components. By combining decoupled optimization steps with soft regularization, the framework achieves stable convergence and enables efficient joint optimization.

4.2. View-specific representation learning

Effective consensus representation learning requires robust and informative view-specific features (Lu et al., 2025). To this end, we first perform view-specific representation learning through two modules: View Reconstruction (VR) and Enhanced View-Specific Representation (EVSR). These modules are designed to preserve intra-view characteristics while incorporating complementary information from other views.

(1) View Reconstruction (VR)

To learn representative embeddings from original views, we design a reconstruction module based on sparse autoencoders. The motivation for choosing sparse autoencoders lies in their ability to suppress redundant information and highlight essential features by enforcing sparsity constraints on hidden representations (Cai et al., 2021). Formally, given the feature X_i^v of the v-th view for region r_i, the reconstruction process is defined as follows:

$$H_i^v = E_v\left(X_i^v\right) \tag{1}$$

$$\widehat{X}_i^v = D_v\left(H_i^v\right) \tag{2}$$

where $E_v(\cdot)$ and $D_v(\cdot)$ denote the encoder and decoder for view v, both implemented as fully connected networks. The hidden representation H_i^v serves as the view-specific representation of the v-th view, while $\widehat{X}_i^v$ denotes the reconstructed feature. The reconstruction loss is:

$$L_r^v = \sum_{r_i \in R} \| X_i^v - \widehat{X}_i^v \|_2^2 \tag{3}$$

To encourage sparsity in the hidden space, we introduce a Kullback-Leibler (KL) divergence constraint (Kullback & Leibler, 1951):

$$L_r = \sum_{v=1}^{V} \omega_r^v L_r^v + \mu \sum_{j=1}^{h} \left(\rho log\left(\frac{\rho}{\widehat{\rho}_j}\right) + (1-\rho) log\left(\frac{1-\rho}{1-\widehat{\rho}_j}\right) \right) \tag{4}$$

where L_r^v and ω_r^v represent the reconstruction loss and its learnable adaptive weight of the view v, respectively. The weight of each view is initialized to $1/V$. The same weighting strategy is applied to the other losses introduced later. μ is a hyperparameter that controls the weight of the sparsity penalty, h is the number of hidden units, ρ is the target sparsity level, and $\widehat{\rho}_j$ is the average activation value of the j-th hidden unit.

(2) Enhanced View-Specific Representation (EVSR)

Initial view-specific features obtained from the VR module are often limited by perspective bias and incomplete information. To address this, each view-specific feature is enhanced by integrating complementary information from other views. Specifically, features from the current view and the remaining views are concatenated along the feature dimension to form a joint representation. This joint representation is then transformed through a nonlinear mapping and combined with the original view-specific feature via element-wise multiplication, resulting in the enhanced representation Z^v:

$$Z^v = H^v \odot M(H) \tag{5}$$

where $\odot$ denotes element-wise multiplication, and $M(\cdot)$ is a nonlinear transformation that projects the joint representation H into the same dimensional space as H^v, thereby ensuring feature-space consistency.

4.3. Consensus representation learning

After obtaining diverse view-specific representations, we integrate them into a unified consensus representation. To achieve this, we introduce the Cross View Transformer (CVTF) module, inspired by Wang et al. (2020). The CVTF is designed to model inter-view interactions and fuse multi-view features into a shared representation. Specifically, feature embeddings Z^v from different views are concatenated and projected into query (Q), key (K), and value (V) spaces using learnable matrices W^Q, W^K, and W^V, respectively. An attention mechanism (Vaswani et al., 2017) is then applied to model pairwise relationships between regions:

$$Z = concat\left(Z^1, Z^2, \ldots, Z^v\right) \tag{6}$$

$$Q = Z \times W^Q, K = Z \times W^K, V = Z \times W^V \tag{7}$$

$$O = softmax\left(\frac{Q \cdot K^\intercal}{\sqrt{d_k}}\right) \tag{8}$$

$$\widehat{H}_i = \sum_{j=1}^{n} O_{ij} \cdot V_j \tag{9}$$

where O denotes the structure-aware similarity matrix. The similarity between regions i and j is measured by the (i,j)-th entry of $Q \cdot K^\intercal$. After softmax normalization, O_{ij} represents the learned structural similarity between region i and region j. Intuitively, regions that exhibit similar semantic and interaction patterns in Z are assigned higher similarity scores O_{ij}. The scaling factor d_k corresponds to the dimensionality of the key vectors and stabilizes the attention computation. To mitigate redundancy introduced by concatenating multiple views, $\widehat{H} = [\widehat{H}_1, \widehat{H}_2, \ldots, \widehat{H}_n]$ is further refined by a fully connected layer, yielding the final consensus representation $\widehat{Z}_c$.

To ensure that the consensus representation effectively captures shared semantic information across views, we introduce a consistency loss that minimizes the discrepancy between the consensus representation and the view-specific representations. The overall consistency loss is computed as:

$$L_{con} = \sum_{v=1}^{V} \omega_{con}^v R(\widehat{Z}_c, Z^v) \tag{10}$$

where $R(\cdot,\cdot)$ denotes the similarity measure between two representations and is implemented using cosine similarity. Minimizing this loss encourages semantic alignment across views while retaining essential view-specific characteristics.

4.4. Multi-view structure-aware contrastive learning

To optimize the learned consensus representation and better align view-specific embeddings, we introduce an innovative contrastive learning approach. Traditional inter-view contrastive methods treat embeddings from different views of the same region as positive pairs, while those from different regions are considered negative samples (Zhang, Long & Cong, 2022). However, this negative sampling strategy suffers from inherent limitations. As discussed in the introduction, treating structurally similar regions as negative samples and forcing their separation during training undermines the model's ability to capture shared structural features. To address this issue, we propose the Multi-View Structure-Aware Contrastive Learning (MVSCL) module, which rrefines the negative sampling strategy by explicitly incorporating structural relationships between regions. The key idea is avoid treating structurally similar regions as negative pairs, thereby avoiding misleading supervision signals.

Specifically, MVSCL leverages a structural similarity matrix O_{ij}, computed by the CVTF module, to quantify the structural similarity between region r_i and regions r_j. Based on this matrix, we design a dynamic weighting mechanism: when O_{ij} is high, indicating strong structural similarity, the term $\exp(-\alpha O_{ij})$ effectively suppresses the contribution of such region pairs as negatives, thereby avoiding misleading supervision signals. Conversely, when O_{ij} is low, the region is treated as a reliable negative sample. In addition, we introduce a dynamic temperature scaling strategy that adjusts the temperature parameter based on structural similarity. This strategy reduces the influence of structurally similar regions during similarity computation, thereby mitigating interference in contrastive learning and promoting more discriminative embeddings. The resulting inter-view contrastive loss L_c^v is defined as follows:

$$L_c^v = \sum_{r_i \in R}\left[-logD\left(\widehat{Z}_{c,i}, Z_i^v\right) + log\left(R\left(\widehat{Z}_{c,i}, Z_i^v\right) + \sum_{r_j \in R} exp\left(-\alpha O_{ij}\right) R'\left(\widehat{Z}_{c,i}, Z_j^v\right)\right)\right] \tag{11}$$

$$R'\left(\widehat{Z}_{c,i}, Z_j^v\right) = exp\left(\frac{\widehat{Z}_{c,i}^\intercal Z_j^v}{\|\widehat{Z}_{c,i}\| \|Z_j^v\| \tau_{ij}}\right) \tag{12}$$

$$\tau_{ij} = \tau\left(1 + \beta O_{ij}\right) \tag{13}$$

where $\widehat{Z}_{c,i}$ is the consensus representation of region r_i. Z_i^v and Z_j^v are the enhanced view-specific representation of region r_i and r_j under view v. τ is the base temperature parameter, and τ_{ij} denotes the adaptive temperature parameters between regions r_i and r_j. α and β represent the dynamic weighting strength and temperature adjustment strength, respectively. $R'(\cdot,\cdot)$ is the quantification function of embedding similarity under temperature control. The total contrastive loss across all views is computed by adaptively weighting and summing the L_c^v of each view:

$$L_c = \sum_{v=1}^{V} \omega_c^v L_c^v \tag{14}$$

4.5. Training strategy based on soft lagrangian constraints

To resolve the gradient conflicts identified in Section 1, we design a coordinated optimization strategy that jointly optimizes feature fusion and representation alignment objectives in a step-wise manner, inspired by the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). Specifically, we propose a multi-stage collaborative optimization framework with soft Lagrangian constraints (Algorithm 1). Unlike the ADMM, which enforces strict equality constraints and updates explicit Lagrange multipliers, our method introduces soft coupling mechanisms that serve as flexible regularization across stages. This design better accommodates the non-convex and multi-objective nature of deep neural networks, while improving training stability and cross-stage coordination. The proposed strategy comprises the following three components:

(1) Multi-stage collaborative optimization strategy: Each training iteration is decomposed into a sequence of sub-tasks: first optimizing the VR module, then jointly optimizing the EVSR and CVTF modules, and finally optimizing the MVSCL module. This order strictly follows the intrinsic information flow of the framework, where the output of one stage serves as the input for the next, reflecting the dependencies among stages. During each stage, only the parameters of the current module are updated, while all others are held fixed. This design reduces gradient interference, enhances training stability, and improves representation quality.
(2) Soft Lagrangian constraint mechanism: To enable flexible coordination across stages, we introduce soft consistency regularization terms between adjacent optimization stages, inspired by classical Lagrangian multiplier methods (Bertsekas, 1997), thereby forming a cross-stage collaborative objective. This mechanism mimics Lagrangian multipliers in a soft form: instead of enforcing exact consistency, it tolerates moderate deviations while penalizing excessive misalignment. The resulting loss is defined as:

$$L_{con} = L_{con} + \lambda_1 \cdot R(Z, H) \tag{15}$$

$$L_c = L_c + \lambda_2 \cdot R(\widehat{Z}_c, \widehat{Z}_m) \tag{16}$$

where $R(\cdot,\cdot)$ measures the similarity between embeddings at adjacent stages using cosine similarity. The hyperparameter λ_1 and λ_2 control the strength of the soft constraints, regulating the cooperation between stages. To avoid disrupting early-stage learning, we employ a delayed activation strategy that activates soft constraints only after a preset number of iterations.

(3) State-aware constraint propagation mechanism: To further improve inter-stage coordination, we introduce a Stage Memory State (SMS) mechanism. In the first two stages, in addition to generating outputs, the model produces a memory state that captures optimization signals. This state is passed to subsequent stages to guide constraint strength. Specifically, the SMS modulates soft constraint strength dynamically:

$$\lambda_1 = \omega \cdot exp(-\varphi \cdot \| \nabla L_r \|_2) \tag{17}$$

$$\lambda_2 = \omega \cdot exp(-\varphi \cdot \| \nabla L_{con} \|_2) \tag{18}$$

where $\| \nabla \boldsymbol{L_r} \|_2$ and $\| \nabla \boldsymbol{L_{con}} \|_2$ represents the norm of the gradient corresponding to the loss function at the first and second stage, ω and φ are hyperparameters controlling the base constraint strength and sensitivity. When the previous stages gradients are large—indicating instability or large updates—the constraint strength λ is reduced to allow more flexible learning and avoid propagating errors. This state-aware mechanism improves model robustness to noise and optimization variance.

5. Experiments

In this section, we conduct a comprehensive evaluation of the proposed SAMC framework by systematically addressing the following key questions:

(Q1) Performance comparison: Does SAMC achieve superior performance compared with state-of-the-art baselines on downstream tasks?
(Q2) Ablation study: Do the individual components of SAMC contribute effectively to overall model performance?
(Q3) Qualitative study: How does the MVSCL module enhance the learned representations, and why does it lead to improved downstream performance?
(Q4) Convergence analysis: Does the proposed multi-stage optimization strategy with soft Lagrangian constraints improve training stability?
(Q5) Complexity analysis: Can SAMC maintain high predictive accuracy while ensuring reasonable computational overhead?
(Q6) Sensitivity analysis: How do variations in hyperparameter settings affect the performance and robustness of SAMC?

Algorithm 1
Training strategy based on Soft Lagrangian constraints.

Input: Multi-view data $X = \{X^1, \ldots, X^V\}$ **Output:**Final node embedding $\widehat{Z}_m$

while not converged **do**
// **Stage 1**
Freeze parameters of EVSR, CVTF, MVSCL
for $v = 1 -> V$ **do**
Compute $H = \{H^1, \ldots, H^V\}$ with Eq. (1)-(2)
Compute reconstruction loss with Eq. (4)
end for
Backpropagate and update parameters of VR
Calculate and Store λ_1 into SMS
// **Stage 2**
Freeze parameters of VR, MVSCL
for $v = 1 -> V$ **do**
Compute $Z = \{Z^1, \ldots, Z^V\}$ and $\widehat{Z}_c$ with Eq. (5)-(9)
Compute consistency loss with Eq. (15)
end for
Backpropagate and update parameters of EVSR, CVTF
Calculate and Store λ_2 into SMS
// **Stage 3**
Freeze parameters of VR,EVSR, CVTF
for $v = 1 -> V$ **do**
Compute contrastive loss with Eq. (16)
end for
Backpropagate and update parameters of MVSCL
end while
return $\widehat{Z}_m$

(Q7) Robustness analysis: Does SAMC exhibit greater robustness than baseline models when data are incomplete? Can SAMC still maintain stable performance in non-function-driven dynamic tasks?
(Q8) Transferability analysis: Does SAMC have the cross-city generalization ability?

5.1. Experimental settings

5.1.1. Datasets

This study used real-world urban datasets collected from the New York City (NYC) Open Data Platform[1] and the Chicago (CHI) Data Portal.[2] Taxi trip records are employed to characterize inter-regional interaction patterns, while Point of Interest (POI) data capture the semantic features of regions. Regional divisions follow administrative boundaries as defined by the U.S. Census Bureau. An overview of the datasets used for model training and downstream task evaluation is provided in Tables 2 and 3, respectively.

5.1.2. Evaluation metrics

Different evaluation metrics are adopted according to the nature of downstream tasks. For regression tasks, such as regional popularity and service demand prediction, we evaluated performance using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). For classification tasks, such as land use classification, we used Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F-measure to provide a comprehensive evaluation of clustering quality.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}(y_i - \widehat{y}_i)^2} \tag{20}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

$$NMI(U, V) = \frac{2 \times I(U; V)}{H(U) + H(V)} \tag{22}$$

[1] https://opendata.cityofnewyork.us/
[2] https://data.cityofchicago.org/

Table 2
Description of datasets used for model training.

Datasets	NYC	CHI
Region	270 spatial blocks delineated by street boundaries in Manhattan	77 spatial blocks delineated by street boundaries in Chicago
POI data	Approximately 10,000 POIs covering 244 attribute types	Approximately 60,000 POIs covering 244 attribute types
Taxi trip records (data collection time)	10 million taxi trip records collected over one month (07/2015 - 08/2015)	3 million taxi trip records collected over one month (03/2021 - 04/2021)

Table 3
Description of datasets used for downstream task evaluation.

Downstream tasks	NYC	CHI
Regional popularity prediction (data collection time)	Check in data: approximately 100,000 records (06/2013 - 11/2014)	Check in data: approximately 160,000 records (06/2013 - 11/2014)
Service demand prediction (data collection time)		Service request data: approximately 25,000 records (01/2022 - 12/2022)
Land use classification	District division: Manhattan is divided into 29 zones based on land use types	
Traffic accident prediction		Traffic accident data: approximately 125,000 records (01/2023 - 12/2023)

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} \tag{23}$$

$$F - measure = 2 \times \frac{Precision}{Precision + Recall} \tag{24}$$

where $y_i, \widehat{y}_i, \overline{y_i}$ denote the truth value, predicted value, and the mean of all truth values for region r_i, respectively. U and V represent the true label partition and the clustering result generated by the model, respectively. $I(U;V)$, $H(U)$ and $H(V)$ refer to the mutual information, the entropy of U, and the entropy of V, respectively. The Rand Index (RI) denotes the number of sample pairs that are correctly grouped together. $E(RI)$ and $max(RI)$ correspond to the expected and maximum values of RI, respectively. *Precision* is the proportion of correctly predicted positive samples among all predicted positive samples, and *Recall* is the proportion of correctly predicted positive samples among all actual positive samples.

5.1.3. Hyperparameter settings

We closely follow the parameter settings recommended in the original studies of the baseline models. In our experiments, the sparsity penalty weight μ is searched within $[10^{-4},10^{-1}]$ and set to 0.01. The reconstruction module uses a three-layer encoder-decoder architecture, with the dimensions of latent representations selected from {64, 128, 256, 512}. The target sparsity level ρ is chosen from {0.05, 0.1, 0.2} and set to 0.1, while the average activation $\widehat{\rho}_j$ is computed dynamically during training. The parameter n corresponds to the total number of regions in each city and is fixed by the dataset. In the base temperature parameter τ for contrastive learning is set to 0.2. Regarding the soft Lagrangian constraints, the balancing coefficients λ_1 and λ_2 are searched within [0.1,2.0] and both set to 1.0 in the final model. For the adaptive constraint components, the base constraint strength ω is selected from [0.1,1.0] and set to 0.5, while the sensitivity parameter φ is selected from [1.0,5.0] and set to 2.0. The Adam optimizer is employed with an initial learning rate of 0.0003 for updating the parameters of EVSR, CVTF, MVSCL module. For the VR module, the learning rate of 0.01 is used. In the experiment, we added an early stopping mechanism to prevent overfitting.

5.1.4. Baseline models

We selected eleven representative models for comparison. To provide a comprehensive and structured comparison, we categorize the baseline methods into four groups according to their core learning paradigms: single-view baseline, feature fusion-based baselines, representation alignment-based baselines and joint learning-based baselines. In the experiment, we ran the implementation code of these baselines on the same dataset and adopted the parameter settings given in the original papers to obtain the experimental results.

(1) Single-view baseline

ZE-Mob (Yao et al., 2018) captures co-occurrence relationships between regions by integrating spatiotemporal features to learn regional embeddings.

(2) Feature fusion-based baselines

MV-PN (Fu et al., 2019) constructs a multi-view network based on POI data and incorporates spatial autocorrelation and Top-k

locality to enhance regional embedding quality.

MVURE (Zhang et al., 2020) improves representation learning by enabling inter-view information sharing and adaptively fusing multi-view representations.

MGFN (Wu et al., 2022) utilizes a multi-level cross-attention mechanism together with a fusion module and a joint learning strategy to obtain informative regional embeddings.

HREP (Zhou et al., 2023) constructs a heterogeneous graph from multi-view data and proposes a relationship-aware graph embedding approach to capture region-level semantics.

ASGCN (Xu et al., 2024) proposes an attention-based framework that separately learns view-specific and consensus embeddings. By incorporating residual connections into the attention mechanism, it effectively fuses complementary information across views while alleviating potential information loss.

(3) Representation alignment-based baselines

ReMVC (Zhang, Long & Cong, 2022) jointly learns representations through both intra-view and inter-view contrastive learning.

MFLVC (Xu et al., 2022) proposes a multi-level feature learning strategy that jointly optimizes low-level feature reconstruction and contrastive consistency of high-level semantic representations, thereby enabling effective multi-view clustering while mitigating the impact of view-specific noise.

ReCP (Li et al., 2024b) combines contrastive learning with reconstruction-based objectives to learn view-specific embeddings. It enhances inter-view consistency by maximizing mutual information across views and minimizing conditional entropy, resulting in more coherent representations.

CureGraph (Li & Zhou, 2025) employs a contrastive multi-view graph representation framework. It models regional spatial dependencies via a spatial autocorrelation matrix while jointly capturing cross-region inter-view interactions and intra-region intra-view relationships to learn regional embeddings.

(4) Joint learning-based baselines

CREME (Zhang et al., 2022b) employs a joint learning framework that combines attention-based fusion with contrastive learning. It maximizes mutual information between the consensus and each view-specific embeddings, while reducing redundancy among view-specific embeddings. This enables more effective capture of cross-view complementary semantics.

5.2. Performance comparison (Q1)

We evaluated the quality of regional embeddings on two representative downstream tasks in urban computing: regional popularity/service demand prediction and land use classification. Following Yang et al. (2014), the number of check-ins and service requests per region were used as proxies for popularity and demand, respectively. For prediction tasks, the embeddings generated by different methods were fed into a Lasso regression model, and performance was assessed using 5-fold cross-validation. For classification, we applied K-means clustering on the learned embeddings to assess their discriminative capacity (Tables 4 and 5).

Our proposed model consistently outperformed baseline methods across all tasks and datasets. For example, in the popularity prediction task on the New York dataset, it achieved a 10.66 % improvement in R^2 over the second-best method, ReCP, demonstrating its strong predictive capability. Similarly, it produced the best clustering results in land use classification, indicating the high quality and discriminative power of the learned representations. Compared with single-view (e.g., ZE-Mob) and simplistic multi-view fusion models (e.g., MV-PN, ReMVC), multi-view learning methods such as MVURE, MGFN, HREP, ReCP, ASGCN, MFLVC, CREME, CureGraph and our approach generally achieve better results by leveraging complementary information from multiple modalities, through contrastive learning, attention-based fusion, or other cross-view consistency mechanisms. However, performance among multi-view

Table 4
Performance comparison of all models on NYC dataset. The optimal and suboptimal performance indicators are marked in bold and underlined.

MODEL	Regional popularity prediction			Land use classification		
	MAE (↓)	RMSE (↓)	R^2 (↑)	NMI (↑)	ARI (↑)	F-measure (↑)
ZE-Mob	263.25	416.65	0.287	0.422	0.060	0.092
MV-PN	276.15	436.31	0.226	0.410	0.042	0.080
MVURE	242.27	370.06	0.515	0.740	0.416	0.431
MGFN	233.64	376.80	0.489	0.749	0.426	0.447
ReMVC	259.65	404.65	0.331	0.764	0.440	0.459
CREME	152.26	276.62	0.715	0.775	<u>0.482</u>	0.498
MFLVC	175.61	304.40	0.684	0.768	0.468	0.486
HREP	150.15	279.52	0.702	<u>0.778</u>	0.480	<u>0.501</u>
ASGCN	186.66	317.61	0.675	0.77	0.476	0.488
ReCP	<u>146.99</u>	<u>260.87</u>	<u>0.722</u>	0.772	0.473	0.492
CureGraph	204.61	324.16	0.654	0.768	0.448	0.476
SAMC	**124.12**	**221.86**	**0.797**	**0.786**	**0.490**	**0.508**

Table 5
Performance comparison of all models on CHI dataset. The optimal and suboptimal performance indicators are marked in bold and underlined.

MODEL	Regional popularity prediction			Service demand prediction		
	MAE (↓)	RMSE (↓)	R^2 (↑)	MAE (↓)	RMSE (↓)	R^2 (↑)
ZE-Mob	1328.62	2686.92	0.259	232.15	325.62	0.153
MV-PN	1036.77	2419.62	0.298	228.3	319.49	0.186
MVURE	911.48	1819.26	0.519	190.62	268.05	0.477
MGFN	893.62	1764.89	0.536	197.48	275.25	0.425
ReMVC	936.12	2216.25	0.496	214.28	299.62	0.296
CREME	<u>512.92</u>	<u>1156.25</u>	<u>0.723</u>	188.71	263.84	0.493
MFLVC	843.64	1714.69	0.569	185.95	263.15	0.495
HREP	769.15	1609.69	0.634	186.61	265.04	0.481
ASGCN	867.58	1739.65	0.543	192.36	271.26	0.465
ReCP	569.65	1359.65	0.709	<u>184.65</u>	<u>260.62</u>	<u>0.509</u>
CureGraph	804.44	1724.65	0.614	189.14	264.45	0.483
SAMC	**455.36**	**914.29**	**0.768**	**176.84**	**247.84**	**0.575**

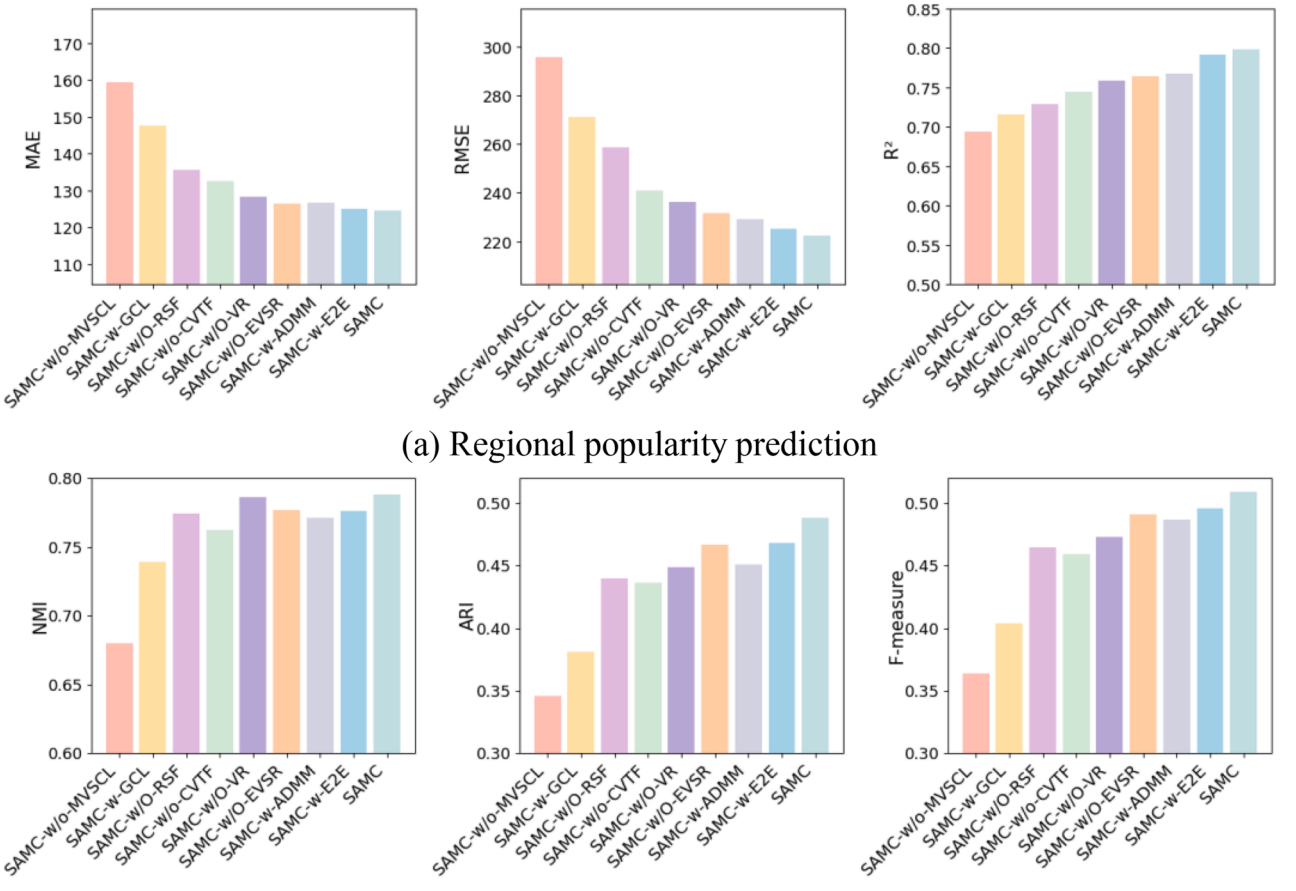
methods still varies depending on the design of fusion and alignment strategies. For example, both CREME and our model adopt a joint learning framework that integrates feature fusion with representation alignment to preserve semantic consistency across views. Our model extends this paradigm by introducing a structure-aware multi-view contrastive learning module, which incorporates structural similarity into negative sampling to enhance semantic coherence and cross-view consistency. In addition, it employs a multi-stage coordinated optimization strategy with soft constraints, which further boosts model robustness and predictive accuracy. These combined innovations produce more reliable and generalizable embeddings, allowing our model to consistently outperform baselines across diverse tasks and evaluation metrics. Compared with the attention-based baselines used in our experiments, SAMC introduces a more expressive mechanism for cross-view interaction. Methods such as ASGCN rely on multi-level or residual attention to fuse view-specific features, while MV-PN incorporates locality-aware attention on POI networks. These approaches primarily operate at the feature level and do not explicitly model global structural relationships among regions. In contrast, the proposed CVTF module adopts a region-level cross-view transformer, which captures long-range structural dependencies both views and across regions, enabling a more comprehensive representation of multi-view spatial patterns. Notably, we further observe that the extent of improvement varies between regression and classification tasks: in regression tasks, SAMC achieved improvements exceeding 15 %, whereas in classification tasks, the improvement mainly ranged from 2 % to 10 %. We further discuss this phenomenon in the discussion section.

5.3. Ablation study (Q2)

To assess the contribution of each component in our framework, we conducted an ablation study on the NYC dataset using four variant models:

(1) **SAMC-w/o-MVSCL:** The MVSCL module is removed.
(2) **SAMC-w-GCL:** The MVSCL module is replaced with a graph contrastive learning module.
(3) **SAMC-w-E2E:** The multi-stage training strategy with soft Lagrangian constraints is replaced by an end-to-end training.
(4) **SAMC-w-ADMM:** The multi-stage training strategy with soft Lagrangian constraints is replaced by ADMM training strategy.
(5) **SAMC-w/o-VR:** The VR module is removed.
(6) **SAMC-w/o-EVSR:** The EVSR module is removed.
(7) **SAMC-w/o-CVTF:** The CVTF module is replaced by simple feature concatenation.
(8) **SAMC-w/o-RSF:** The regional semantic features view is removed.

As shown in Fig. 2, our proposed model consistently outperformed all eight variants, demonstrating the critical role of each component. Among these, removing the MVSCL module (SAMC-w/o-MVSCL) results in the most significant performance degradation, particularly in the land use classification task. This highlights the pivotal role of MVSCL in enhancing representation quality. By incorporating structural similarities into the negative sampling process, the MVSCL module reduces the risk that semantically or functionally similar regions are erroneously pushed apart in the embedding space. Such capability is critical for capturing fine-grained regional distinctions and pmaintaining semantic coherence across views. To further examine MVSCL's contribution, we conduct an additional ablation experiment, denoted as SAMC-w-GCL, in which the MVSCL module is replaced with a graph contrastive learning module. The results show that SAMC consistently outperforms SAMC-w-GCL. GCL typically defines positive pairs through graph neighborhoods, which are determined by predefined edges, such as spatial adjacency. While effective in general graph domains, this strategy rigidly treats all non-neighboring regions as negative samples. As a result, functionally similar but non-neighboring regions may be misclassified, thereby weakening semantic coherence. Unlike GCL, our structure-aware contrastive learning does not rely on a fixed graph structure. Instead, it dynamically constructs structural relationships by fusing multi-view urban signals (e.g., dynamic interaction flows and POI semantic information). MVSCL employs the CVTF module to learn a dynamic structural similarity matrix from multi-view features and introduces adaptive weighting and temperature mechanisms to refine the contrastive process: functionally similar regions are prevented from being improperly separated, while distinctions between functionally dissimilar regions are



(a) Regional popularity prediction

(b) Land use classification

Fig. 2. Performance comparison of model variants on the NYC dataset.

explicitly reinforced. In essence, MVSCL more accurately captures the inherently functional and dynamic nature of urban regional relationships, rather than relying on the static spatial topology assumed by GCL. This design enables SAMC to produce embeddings that are more reliable and better aligned with real-world urban tasks, such as regional popularity prediction and service demand forecasting. The resulting performance degradation confirms that structure-aware negative sampling is essential for avoiding misleading optimization signals and for preserving cross-view semantic consistency.

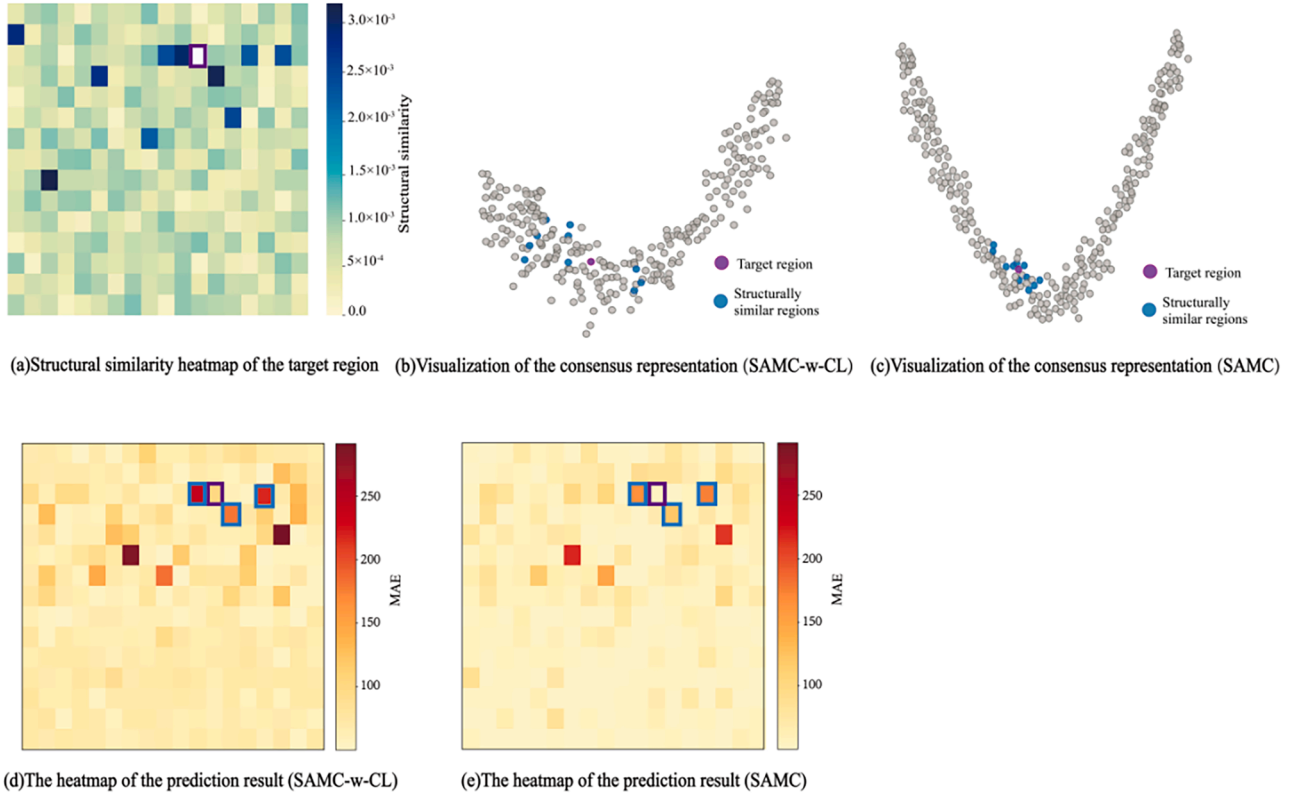
In addition, the effectiveness of the proposed multi-stage coordinated optimization strategy is evidenced by a 1.95 % to 2.85 % and a 2.73 % to 3.69 % improvement in classification accuracy over the end-to-end variant (SAMC-w-E2E) and the training strategy based on ADMM (SAMC-w-ADMM). This performance gain can be attributed to the strategy's ability to alleviate parameter conflicts through stage-wise decoupling and to promote stable convergence via soft constraint regularization. Lastly, removing either the VR or EVSR module leads to a noticeable performance drop, confirming that VR effectively preserves intra-view structural information, while EVSR enhances both view-specific and complementary cross-view representations. This highlights the importance of their joint design in achieving the strong performance of SAMC. Replacing the CVTF module with simple feature concatenation leads to substantial performance degradation, indicating the critical role of CVTF in capturing cross-view functional dependencies among regions. To validate the necessity of using multiple input views within the CVTF module, we conduct an ablation study by removing the regional semantic view (e.g., POI features). The results show a significant performance drop, as this view provides essential functional cues that are critical for learning meaningful inter-regional similarities. Since the structural similarity matrix O is derived end-to-end solely from the multi-view representation Z, removing the semantic view alters Z and consequently degrades the quality of O. This demonstrates that O adaptively integrates complementary urban signals, such as functional semantics and interaction patterns, within a unified, learnable framework.

5.4. Qualitative study (Q3)

To evaluate the effectiveness of the proposed structure-aware contrastive learning approach, we compared the full SAMC model with a variant, denoted as SAMC-w-CL, in which the structure-aware module is replaced by a standard contrastive learning component. We first calculated the structural similarity matrix for the target region (the purple box in Fig. 3(a)) and visualized it as a heatmap, where higher similarity scores indicate stronger structural resemblance. Based on this matrix, we identified the top 10 structurally similar regions and visualized their embedding distributions using t-SNE. As shown in Fig. 3(c), SAMC preserves the proximity of structurally similar regions in the embedding space, indicating its ability to retain structural consistency. In contrast, SAMC-w-CL (Fig. 3(b)) fails to cluster these regions together, as it lacks structure-aware sampling and consequently misidentifies similar regions as negative samples. This flaw impairs representation learning and degrades predictive performance. To further support this observation, Figs. 3(d) and 3(e) display the prediction accuracy heatmaps produced by SAMC-w-CL and SAMC, respectively. The target region and several of its structurally similar neighbors are highlighted with purple and blue boxes. Compared to Fig. 3(d), the SAMC (Fig. 3(e)) achieves notably higher prediction accuracy in both the target region and its structurally similar areas. This improvement can be attributed to SAMC's dynamic incorporation of structural similarity into the negative sampling process, which mitigates the negative impact of semantically misleading supervision and yields more robust prediction results.

To investigate how the CVTF and MVSCL modules contribute to preserving structural consistency and semantic clustering in SAMC, we performed a regional-level study by visualizing embeddings at different stages using t-SNE and selecting a representative region to examine cluster compactness and dispersion throughout the learning process. Fig. 4(b) and 4(c) show the embeddings of the regional semantic and interaction views, while Fig. 4(d) presents the simple concatenation of multi-view features. We selected the target region marked by the red box in Fig. 4(b) (denoted as Region a). After feature concatenation (Fig. 4(d)), regions belonging to the same class as the target (shown as enlarged dark blue points) exhibit a more linear arrangement. However, their overall dispersion remains relatively large, indicating limited improvement in cluster compactness. In contrast, the CVTF module effectively integrates the regional semantic and interaction views, aligning regions with similar functional and structural patterns into tighter clusters (Fig. 4(e)). Compared to the simple concatenation of the two input views (Fig. 4(d)), the CVTF embeddings exhibit significantly improved cluster compactness. To further analyze the internal behavior of SAMC, we examined a misclassified region (highlighted in orange in Fig. 4(b) and labeled as Region b) to understand how SAMC handles structurally and functionally similar regions. For this misclassified region, the negative-sample weight distribution in MVSCL (Fig. 4(g)) shows that structurally similar regions are assigned very low contrastive weights. This explains why these regions are not pushed apart during optimization. By adaptively reducing repulsion between regions sharing strong structural similarity, MVSCL preserves semantic continuity in the embedding space. This micro-level analysis provides clear and interpretable evidence of how CVTF enhances cross-view consistency and how MVSCL governs the contrastive learning process, offering deeper insight into SAMC's behavior at the regional level.

Building upon this micro-level understanding, we further examined the embeddings learned at different training stages from a global perspective to analyze the effectiveness of the proposed model in improving classification accuracy. As shown in Fig. 5(b) to 5(f), compared to the view-specific representations (Fig. 5(b) and 5(c)) and the consensus representation learned by the CVTF module (Fig. 5(d)), the final enhanced consensus representation (Fig. 5(f)) exhibits more compact intra-class clusters and clearer inter-class separations in the embedding space. In contrast, the embedding generated by the state-of-the-art model HREP (Fig. 5(a)) shows less cohesive clustering and more ambiguous class boundaries. Notably, several regions belonging to the same category are dispersed across multiple clusters (highlighted by green circles), indicating inconsistent representation learning. These findings indicate that our model is capable of learning more robust and discriminative representations, which in turn lead to substantial improvements in classification performance.



(a)Structural similarity heatmap of the target region (b)Visualization of the consensus representation (SAMC-w-CL) (c)Visualization of the consensus representation (SAMC)

(d)The heatmap of the prediction result (SAMC-w-CL) (e)The heatmap of the prediction result (SAMC)

Fig. 3. Comparative of embedding structures and prediction accuracy for regional popularity prediction on the NYC dataset: SAMC vs. SAMC-w-CL.

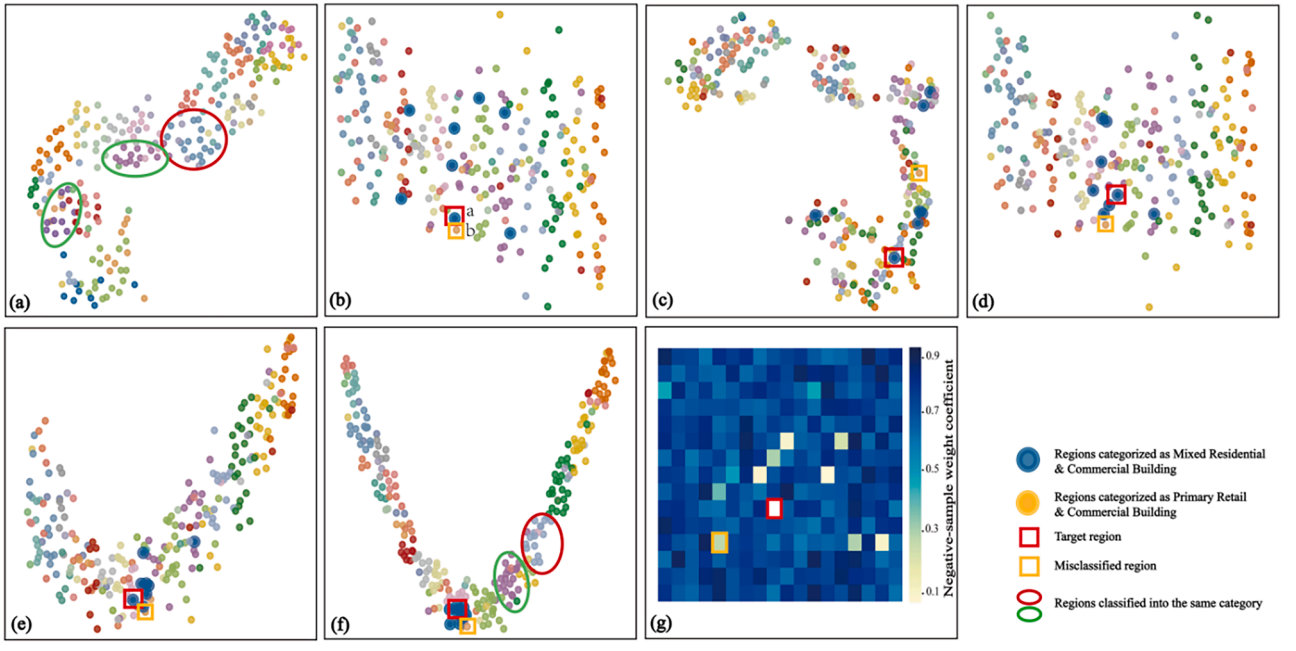
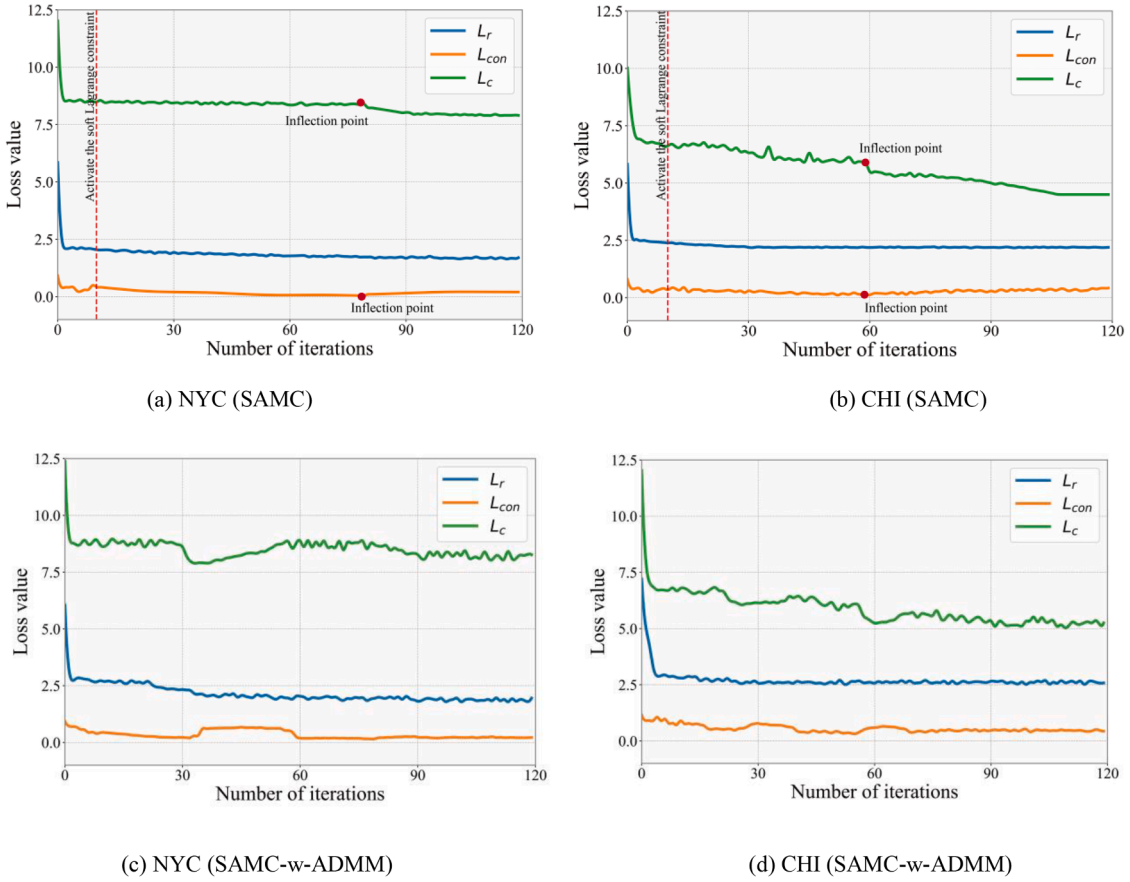


Fig. 4. Visualization of regional embedding representations. Points with the same color represent regions of the same land use type. Points corresponding to regions of the same type as the target region are enlarged to better observe their aggregation. Red boxes indicate the target region, while orange boxes indicate region that were incorrectly classified as belonging to the same category as the target region. (a) HREP; (b) SAMC (VR, regional semantic view); (c) SAMC (VR, regional interaction view); (d) SAMC (Concatenated multi-view); (e) SAMC (CVTF); (f) SAMC (MVSCL); (g) negative-sample weight coefficient of the target region.



(a) NYC (SAMC) (b) CHI (SAMC)

(c) NYC (SAMC-w-ADMM) (d) CHI (SAMC-w-ADMM)

Fig. 5. Convergence curves of the three loss components. The red dashed line indicates the activation point of the soft Lagrangian constraint.

5.5. Convergence analysis (Q4)

Fig. 5 presents the convergence trajectories of model's loss components under different training strategies, clearly highlighting the differences between SAMC and SAMC-w-ADMM. For the ADMM-based training strategy (Fig. 5(c) and 5(d)), the use of strict equality constraints and explicit updates of the Lagrangian multipliers introduces strong conflicts among the loss components, leading to pronounced fluctuations in L_{con} and L_c, and resulting in an overall unstable downward trend. This indicates that the hard constraint mechanism of ADMM tends to exacerbate gradient conflicts and weaken convergence stability. In contrast, the proposed multi-stage soft Lagrangian optimization strategy exhibits a smoother and more coordinated convergence process (Fig. 5(a) and 5(b)). First, the reconstruction loss L_r, consistency loss L_{con} and contrastive loss L_c all exhibit a rapid decline during the initial training phases. This reflects the effectiveness of the proposed multi-stage optimization strategy, which decouples parameter updates across modules to alleviate gradient conflict and promote more stable convergence. Second, L_{con} shows slight fluctuations at the beginning, which are effectively mitigated after the activation of the soft Lagrangian constraint (the red dashed line in Fig. 5(a) and 5(b)). This adjustment results in a smoother downward trend. Third, as L_{con} stabilizes, the contrastive loss L_c associated with the MVSCL module also begins to decline steadily. This suggests that the formation of high-quality consensus representations lays a solid foundation for subsequent structure-aware contrastive learning. Notably, L_c shows an accelerated decrease in the later training stages, with its inflection point closely aligning with a slight uptick in L_{con}. This phenomenon indicates that the cross-stage coupling mechanism dynamically balances the objectives of fusion and alignment, further refining the learned representations. Overall, although slight oscillations are observed when loss values approach convergence, the training process remains stable and well-coordinated. These results validate the effectiveness of the proposed soft Lagrangian multi-stage training strategy in enhancing optimization stability and promoting inter-module collaboration.

5.6. Complexity analysis (Q5)

To evaluate the computational complexity of the SAMC model, we compared it with nine state-of-the-art deep learning models, as shown in Fig. 6. In terms of runtime, SAMC exhibits a moderate execution time, being slightly slower than lightweight models such as ZE-Mob and MV-PN, but significantly faster than more complex architectures, including MGFN and ReCP. This observation suggests

that SAMC strikes a favorable balance between computational efficiency and predictive capability. Regarding parameter size, SAMC contains substantially fewer parameters than complex models such as MGFN and MVURE, enhancing its feasibility for deployment in resource-constrained environments. Although it involves more parameters than lightweight models, SAMC consistently delivers superior prediction accuracy. This highlights its effectiveness in balancing model complexity with performance. Overall, SAMC achieves a favorable compromise between computational cost and performance. Its moderate runtime and compact parameterization underscore its scalability and applicability in real-world settings that demand both efficiency and accuracy.

5.7. Sensitivity analysis (Q6)

To investigate the influence of the hyperparameters α and β in the MVSCL module on model performance, we conducted a detailed sensitivity analysis. As shown in Fig. 7(a) and 7(b), we adopted two representative metrics: Normalized Mutual Information (NMI) for land use classification and Mean Absolute Error (MAE) for regional popularity prediction on the NYC dataset. The results reveal a clear trend: performance improves with increasing α, peaking at $\alpha = 5$, after which a slight decline is observed. The best performance is achieved when $\alpha = 5$ and $\beta = 1.5$. These findings indicate that suboptimal settings of α or β can adversely affect model effectiveness. Specifically, a small α may underemphasize structural similarity, limiting the suppression of false negatives, while an excessively large α may filter out informative contrastive signals. Similarly, an excessively small β may result in insufficient temperature scaling, resulting in structurally similar regions still being regarded as strong negative samples, thereby increasing the risk of false penalties. In contrast, an overly large β weakens the contrastive strength, making it harder to separate dissimilar regions in the embedding space. Overall, these results highlight the importance of appropriately calibrating α and β to balance structural awareness and contrastive strength, thereby improving the robustness and quality of the learned representations. To better assess sensitivity, we also conducted the above experiments on the CHI dataset (Fig. 7(c) and 7(d)). Combined with the results on the NYC dataset, these findings show that SAMC exhibits robust performance across a moderate range: $\alpha \in [3, 5]$ and $\beta \in [0.5, 2]$. Within this range, performance variations are minor, and the optimal configuration consistently falls within the same interval across datasets. These results demonstrate that SAMC is not overly sensitive to these hyperparameters.

5.8. Robustness analysis (Q7)

To validate the robustness of SAMC, we conducted experiments under scenarios simulating data quality imbalances, such as missing POI categories and sparse taxi trajectory data. For comparison, we selected strong baseline models, including CREME, HREP, CureGraph, and MVURE, which are known for their competitive performance in urban representation learning tasks. In the NYC dataset, we randomly removed specific POI categories (e.g., open space and outdoor recreation-related POIs) and evaluated the impact on land use classification. The results show that SAMC exhibited the smallest performance degradation among all methods: NMI for SAMC decreased by 2.7 %, while CREME dropped by 6.1 % and CureGraph by 8.5 % (Table 6). Similarly, in the CHI dataset, we removed taxi trajectories with destinations in 10 % of randomly selected regions and tested service demand prediction. MAE for SAMC increased by only 3.4 %, compared to 7.2 % for HREP and 5.9 % for ReCP (Table 6). These results indicate that SAMC maintains superior robustness under incomplete data. The robustness can be attributed to its coordinated fusion and alignment mechanisms. Through multi-view fusion, SAMC integrates information from multiple data sources, allowing it to compensate for missing or noisy data in a view or a structurally similar region using information from other views or regions. The alignment strategy ensures that the consensus representation remains informative even when individual views are noisy or incomplete.

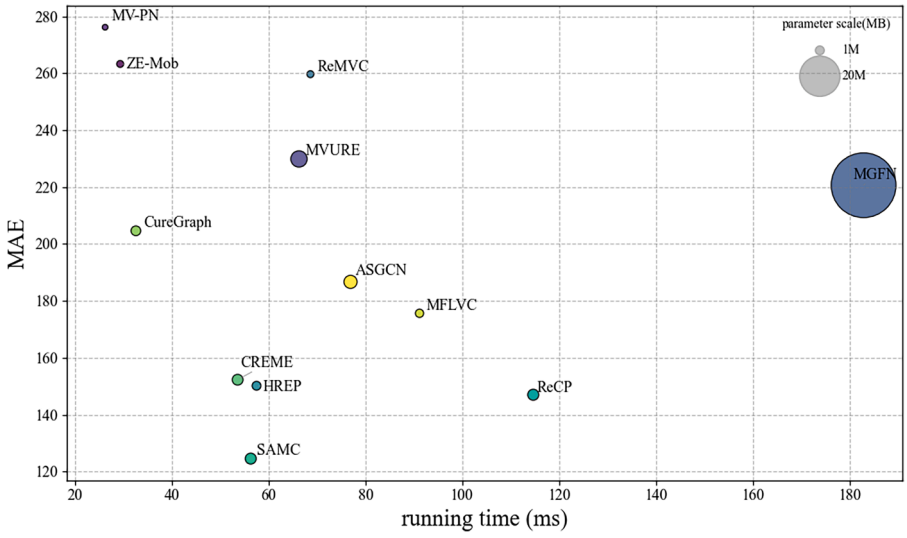


Fig. 6. Results of model complexity comparison.

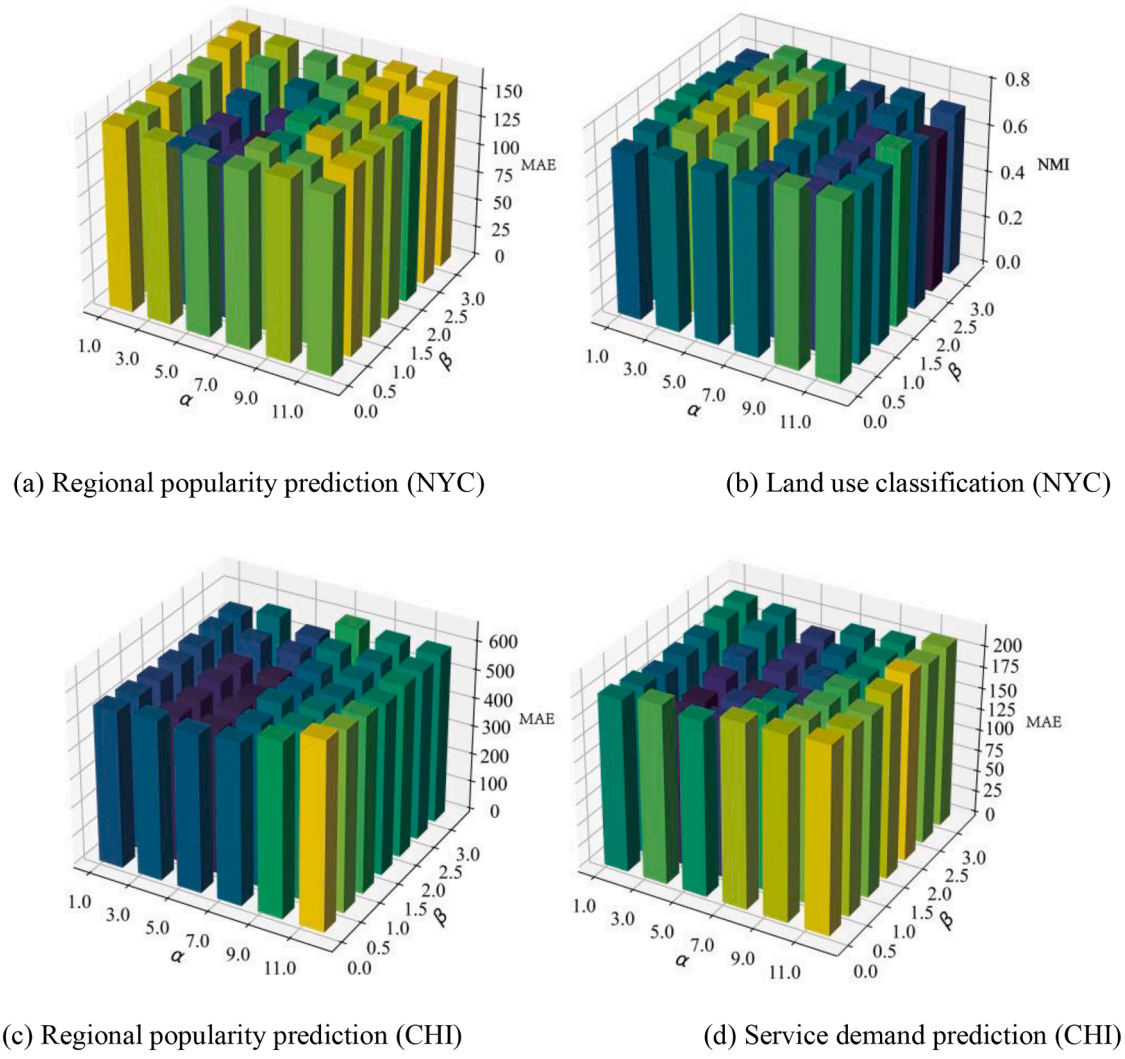


Fig. 7. Results of hyperparameter sensitivity analysis on the NYC dataset.

Table 6
Comparison of model robustness to incomplete data. The decline rate is indicated in parentheses.

MODEL	Land use classification (NYC)			Service demand prediction (CHI)		
	NMI (↑)	ARI (↑)	F-measure (↑)	MAE (↓)	RMSE (↓)	R^2 (↑)
CREME	0.727 (6.1 %)	0.455 (5.7 %)	0.469 (5.7 %)	199.46 (6.9 %)	278.87 (7.8 %)	0.465 (8.2 %)
HREP	0.741 (4.8 %)	0.456 (4.9 %)	0.478 (4.4 %)	200.05 (7.2 %)	276.70 (7.6 %)	0.460 (8.1 %)
ReCP	0.728 (5.7 %)	0.448 (5.2 %)	0.463 (5.8 %)	195.54 (5.9 %)	275.74 (5.3 %)	0.479 (6.3 %)
CureGraph	0.703 (8.5 %)	0.412 (8.1 %)	0.436 (8.3 %)	199.16 (5.3 %)	286.40 (5.9 %)	0.443 (5.9 %)
SAMC	0.765 (2.7 %)	0.476 (2.8 %)	0.493 (2.9 %)	182.85 (3.4 %)	255.03 (3.6 %)	0.558 (3.4 %)

To further evaluate the robustness and generality of SAMC, we conducted an experiment on a non–function-driven dynamic task, namely traffic accident prediction. We formulate the task as a binary classification problem that predicts whether a region will experience at least one accident on the following day. The input features include temporal features, historical accident statistics, and the regional embeddings learned by different models. Specifically, the baseline LSTM uses only temporal features and historical accident statistics as inputs. For embedding-based methods including CREME+LSTM and SAMC+LSTM, we augment these inputs with the learned region embeddings, enabling the LSTM to leverage both temporal dynamics and spatial representations for prediction. It is important to note that we did not employ complex sequence models, since the goal of this experiment is not to pursue the best possible accident prediction performance, but rather to examine the contribution and robustness of the SAMC-learned representations in a dynamic setting. As shown in Table 7, SAMC+LSTM achieves the best performance and significantly outperforms all other baselines.

This demonstrates that the region representations learned by SAMC encode rich and transferable urban knowledge that remains effective even in downstream tasks where functional similarity plays a minor role. This effect can be explained from two perspectives. First, CVTF captures cross-regional human mobility patterns and structural similarities through multi-view feature learning, indirectly reflecting similarities in traffic volume and exposure risk among regions. Second, the structure-aware contrastive learning in MVSCL prevents structurally similar regions from being treated as strong negative samples, thereby preserving correlations among regions with similar risks. Regions with similar structural and semantic tend to exhibit similar patterns of traffic accidents (Chen et al., 2024). Our method leverages this by capturing structural similarities between regions, thereby enhancing prediction performance. Overall, these results indicate that the representations learned by SAMC are not only effective for function-related tasks but can also generalize to dynamic, multi-factor urban prediction scenarios, thereby confirming the robustness and broad applicability of the proposed framework.

5.9. Transferability analysis (Q8)

To investigate the transferability of SAMC across cities, we conducted a Leave-One-City-Out (LOCO) evaluation between NYC and CHI datasets. Specifically, the SAMC encoder was trained on NYC, a large metropolis with 270 spatial blocks, and then transferred to CHI, a mid-sized city with 77 spatial blocks. Given these differences, we adopted a few-shot fine-tuning protocol, where the NYC-pretrained model was fine-tuned using only 25 % of CHI's training data before evaluation on CHI's target tasks. For reference, we also report the performance of a model fully trained on CHI, referred to as the "Oracle". As shown in Table 8, after 25 % fine-tuning, the transferred model's performance decreased by only 10.29 % for regional popularity prediction and 13.79 % for service demand prediction compared with the Oracle. This indicates that SAMC can effectively transfer knowledge across cities of different scales and adapt to new urban contexts with limited supervision, suggesting robustness to variations in both city structure and data volume.

6. DISCUSSION

Joint learning methods that combine feature fusion with contrastive learning have shown substantial potential in urban representation learning. However, they often ignore hidden structural relationships among regions when selecting negative samples for contrastive learning. This oversight can create inconsistent supervision, which reduces both the quality and discriminative power of the learned embeddings. In contrast, the proposed SAMC framework addresses this limitation by introducing a structure-aware contrastive learning mechanism that explicitly incorporates functional similarity into the negative sampling process. This mechanism not only preserves semantic consistency but also improves the discriminability of the learned representations. Furthermore, ablation experiments confirm the effectiveness of this module, showing that SAMC explicitly captures structural aspects that existing joint learning methods often ignore. Another important distinction lies in the optimization strategy. Most conventional joint learning approaches rely on end-to-end optimization of multiple objectives, which can induce gradient conflicts and lead to unstable convergence. By contrast, SAMC adopts a coordinated optimization scheme based on soft Lagrangian constraints, which dynamically mediates gradient interactions between the fusion and alignment modules. This adaptive mechanism alleviates gradient interference, resulting in more stable convergence and improved performance across multiple modules.

From a theoretical perspective, the soft Lagrangian optimization framework advances multi-module and multi-view representation learning by providing a principled and generalizable approach for handling competing objectives. Although originally developed for urban representation learning, this framework is also applicable to other domains. For instance, in multi-modal emotion recognition systems such as CARAT (Peng et al., 2024), reconstruction objectives can conflict with contrastive alignment objectives. Similarly, in recommendation systems like DSL (Wang et al., 2023), supervised and self-supervised objectives may exert opposing influences when learning user representations. In both cases, soft Lagrangian constraints provide a dynamic mechanism for balancing these conflicts, thereby improving training stability and generalization.

From a practical perspective, SAMC demonstrates strong applicability across a range of urban analytics tasks, including functional area identification, urban planning, and transportation management. By capturing meaningful structural patterns among regions, the learned embeddings provide reliable inputs for downstream tasks such as land-use classification, regional clustering, and demand forecasting. Moreover, SAMC achieves these benefits with relatively low computational cost and a compact model size, enabling efficient training and inference. This efficiency supports scalable deployment in real-world urban scenarios, as demonstrated by the cross-city transfer experiments in Section 5.9. Overall, the combination of strong predictive performance and computational efficiency

Table 7
Performance comparison in traffic accident prediction on the CHI dataset.

MODEL	Traffic accident prediction (CHI)		
	NMI (↑)	ARI (↑)	F-measure (↑)
LSTM	0.645	0.489	0.514
CREME+LSTM	0.659	0.497	0.520
HREP+LSTM	0.964	0.521	0.558
ReCP+LSTM	0.664	0.501	0.531
CureGraph+LSTM	0.576	0.446	0.486
SAMC+LSTM	0.728	0.532	0.595

Table 8
Cross-city transferability evaluation between NYC and CHI (LOCO setting).

MODEL	Regional popularity prediction			Service demand prediction		
	MAE (↓)	RMSE (↓)	R^2 (↑)	MAE (↓)	RMSE (↓)	R^2 (↑)
SAMC (Oracle, CHI-trained)	**455.36**	**914.29**	**0.768**	**176.84**	**247.84**	**0.575**
SAMC (25 % Fine-tuning)	521.63	1165.67	0.717	188.54	276.51	0.503
Avg. Drop (%)	13.79			10.29		

highlights the practical value and broad applicability of SAMC.

To further understand SAMC's practical implications, we analyzed its performance on different task types, specifically regression and classification. Regression tasks, such as predicting regional popularity or service demand, benefit more substantially from SAMC than classification tasks. This difference arises from the distinct learning characteristics of the two task types. Regression requires modeling smooth and continuous input–output relationships, which are highly sensitive to the structural continuity of learned representations. SAMC explicitly reinforces these structural properties through two complementary mechanisms: the Cross-View Transformer (CVTF), which captures high-order structural dependencies across regions and views using attention-based relational modeling, and the Multi-View Structure-Aware Contrastive Learning (MVSCL) module, which maintains structural consistency by preventing structurally similar regions from being separated during optimization. Together, these mechanisms encourage a representation space with smooth transitions among structurally related regions, enabling the model to capture gradual variations essential for accurate regression predictions. In contrast, classification tasks typically rely on clearly separated functional categories and emphasize categorical discrimination; therefore, the structural consistency imposed by SAMC provides comparatively limited benefits.

7. CONCLUSIONS

This paper introduced SAMC, a structure-aware multi-view representation learning framework designed to enhance structural awareness and alleviate optimization conflicts between fusion and contrastive learning objectives. To address the issue of suboptimal negative sampling, SAMC incorporates a structure-guided contrastive module that leverages latent structural similarity to avoid treating semantically similar regions as negatives, thereby improving semantic coherence and cross-view consistency. Additionally, to resolve the inherent optimization conflict between feature fusion and alignment, SAMC adopts a multi-stage training strategy that integrates soft Lagrangian constraints to coordinate gradient updates across modules. By decoupling the learning process and applying soft regularization to intermediate outputs, SAMC stabilizes convergence and improves training efficiency. Extensive experiments on multiple real-world urban datasets demonstrate that SAMC consistently outperforms state-of-the-art baselines across a variety of downstream tasks. These results confirm that SAMC not only enhances the discriminability and robustness of learned embeddings but also achieves greater training stability through coordinated optimization.

Despite its promising performance, SAMC has certain limitations. First, current implementations primarily rely on pairwise structural relationships between regions, limiting their capacity to capture higher-order interactions in complex urban systems. A promising direction for future research is to incorporate hypergraph neural networks, which model region relationships at the hyperedge level and better capture intricate spatial dependencies. Second, while SAMC demonstrates robustness to hyperparameter variations, its currently involves a degree of manual configuration in hyperparameter selection and the design of the stage-wise training protocol. Future work could explore automated approaches, such as Bayesian optimization for hyperparameter tuning or learnable stage scheduling, to further improve reproducibility and ease of deployment.

CRediT authorship contribution statement

Jinghui Wei: Writing – original draft, Visualization, Methodology, Formal analysis. **Sheng Wu:** Writing – review & editing, Project administration, Funding acquisition. **Shifen Cheng:** Writing – review & editing, Supervision, Conceptualization. **Peixiao Wang:** Formal analysis. **Feng Lu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by National key Research and Development Program of China project (No. 2023YFB3906804).

Data availability

Data will be made available on request.

References

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society, 48*(3), 334. -334.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning, 3*(1), 1–122.

Cai, J., Wang, S., & Guo, W. (2021). Unsupervised embedded feature learning for deep clustering with stacked sparse auto-encoder. *Expert Systems with Applications, 186*, Article 115729. Article.

Cao, J., Wang, X., Chen, G., Tu, W., Shen, X., Zhao, T., … Li, Q. (2025a). Disentangling the hourly dynamics of mixed urban function: A multimodal fusion perspective using dynamic graphs. *Information Fusion, 117*, Article 102832. Article.

Cao, J., Wang, X., Chen, J., Tu, W., Li, Z., Yang, X., … Li, Q. (2025b). Urban representation learning for fine-grained economic mapping: A semi-supervised graph-based approach. *ISPRS Journal of Photogrammetry and Remote Sensing, 226*, 317–331.

Chan, W., & Ren, Q. (2023). Region-wise attentive multi-view representation learning for urban region embedding. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 3763–3767).

Chen, M., Yuan, H., Jiang, N., Bao, Z., & Wang, S. (2024). Urban traffic accident risk prediction revisited: Regionality, proximity, similarity and sparsity. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 281–290).

Chen, Y., Huang, W., Zhao, K., Jiang, Y., & Cong, G. (2025). Self-supervised representation learning for geospatial objects: A survey. *Information Fusion*, Article 103265. Article.

Deng, M., Chen, C., Zhang, W., Zhao, J., Yang, W., Guo, S., … Luo, J. (2024). HyperRegion: Integrating graph and hypergraph contrastive learning for region embeddings. *IEEE Transactions on Mobile Computing, 5*(24), 3667–3684.

Fang, S., Pan, X., Xiang, S., & Pan, C. (2021). Meta-MSNet: Meta-learning based multi-source data fusion for traffic flow prediction. *IEEE Signal Processing Letters, 28*, 6–10.

Fu, H., Wei, Y., Chen, G., He, X., Gao, Q., & Zhou, F. (2025). Augmented graph information bottleneck with type-aware periodicity heterogeneity for explainable crime prediction. *Information Processing & Management, 62*(6), Article 104227. Article.

Fu, Y., Wang, P., Du, J., Wu, L., & Li, X. (2019). Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 906–913.

Han, X., Jiang, Z., Liu, N., Song, Q., Li, J., & Hu, X. (2022). Geometric graph representation learning via maximizing rate reduction. In *Proceedings of the ACM Web Conference 2022* (pp. 1226–1237).

Huang, W., Zhang, D., Mai, G., Guo, X., & Cui, L. (2023). Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing, 196*, 134–145.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., & Ermon, S. (2019). Tile2vec: Unsupervised representation learning for spatially distributed data. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 3967–3974.

Jin, J., Song, Y., Kan, D., Zhang, B., Lyu, Y., Zhang, J., & Lu, H. (2024). Learning context-aware region similarity with effective spatial normalization over Point-of-Interest data. *Information Processing & Management, 61*(3), Article 103673. Article.

Ke, G., Chao, G., Wang, X., Xu, C., Zhu, Y., & Yu, Y. (2023). A clustering-guided contrastive fusion for multi-view representation learning. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(4), 2056–2069.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79–86.

Li, J., Zhou, C., Ji, X., Li, M., Lu, G., Xu, Y., & Zhang, D. (2024a). Multi-view instance attention fusion network for classification. *Information Fusion, 101*, Article 101974. Article.

Li, J., & Zhou, X. (2025). CureGraph: Contrastive multi-modal graph representation learning for urban living circle health profiling and prediction. *Artificial Intelligence, 340*, Article 104278.

Li, Y., Yang, M., & Zhang, Z. (2018). A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering, 31*(10), 1863–1883.

Li, Y., Huang, W., Cong, G., Wang, H., & Wang, Z. (2023). Urban region representation learning with OpenStreetMap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1363–1373).

Li, Z., Huang, C., Xia, L., Xu, Y., & Pei, J. (2022). Spatial-temporal hypergraph self-supervised learning for crime prediction. In *2022 IEEE 38th International Conference on Data Engineering* (pp. 2984–2996).

Li, Z., Huang, W., Zhao, K., Yang, M., Gong, Y., & Chen, M. (2024b). Urban region embedding via multi-view contrastive prediction. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(8), 8724–8732.

Liu, S., Kong, Z., Huang, T., Du, Y., & Xiang, W. (2024). An ADMM-LSTM framework for short-term load forecasting. *Neural Networks, 173*, Article 106150. Article.

Lu, M., Zhang, Q., & Chen, B. (2025). Divergence-guided disentanglement of view-common and view-unique representations for multi-view data. *Information Fusion, 114*, Article 102661. Article.

Luo, Y., Chung, F. L., & Chen, K. (2022). Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 4294–4298).

Ma, A., Yu, Y., Shi, C., Guo, Z., & Chua, T. S. (2024). Cross-view hypergraph contrastive learning for attribute-aware recommendation. *Information Processing & Management, 61*(4), Article 103701. Article.

Neal, I., Seth, S., Watmough, G., & Diallo, M. S. (2022). Census-independent population estimation using representation learning. *Scientific Reports, 12*(1), 5185. Article.

Peng, C., Chen, K., Shou, L., & Chen, G. (2024). CARAT: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition. In *, 38. Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 14581–14589).

Robinson, C., Hohman, F., & Dilkina, B. (2017). A deep learning approach for population estimation from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities* (pp. 47–54).

Sun, F., Qi, J., Chang, Y., Fan, X., Karunasekera, S., & Tanin, E. (2024). Urban region representation learning with attentive fusion. In *2024 IEEE 40th International Conference on Data Engineering* (pp. 4409–4421).

Trosten, D. J., Lokse, S., Jenssen, R., & Kampffmeyer, M. (2021). Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1255–1265).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (p. 30).

Vomfell, L., Härdle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems, 113*, 73–85.

Wang, H., & Li, Z. (2017). Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 815–822).

Wang, H., Zhang, W., Wang, Q., & Ma, X. (2025). Adaptive structural-guided multi-level representation learning with graph contrastive for incomplete multi-view clustering. *Information Fusion, 119*, Article 103035. Article.

Wang, J., Huang, W., & Biljecki, F. (2024). Learning visual features from figure-ground maps for urban morphology discovery. *Computers, Environment and Urban Systems,, 109*, Article 102076. Article.

Wang, P., Fu, Y., Zhang, J., Li, X., & Lin, D. (2018). Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology, 9*(6), 1–28.

Wang, R., Li, L., Tao, X., Wang, P., & Liu, P. (2022a). Contrastive and attentive graph learning for multi-view clustering. *Information Processing & Management, 59*(4), Article 102967. Article.

Wang, S., Lin, X., Fang, Z., Du, S., & Xiao, G. (2022b). Contrastive consensus graph learning for multi-view clustering. *IEEE/CAA Journal of Automatica Sinica, 9*(11), 2027–2030.

Wang, T., Xia, L., & Huang, C. (2023). Denoised self-augmented learning for social recommendation. *arXiv preprint* arXiv:2305.12685.
Wang, Z., Li, H., & Rajagopal, R. (2020). Urban2vec: Incorporating street view imagery and POIs for multi-modal urban neighborhood embedding. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(1), 1013–1020.
Wu, S., Yan, X., Fan, X., Pan, S., Zhu, S., Zheng, C., Cheng, M., & Wang, C. (2022). Multi-graph fusion networks for urban region embedding. arXiv preprint arXiv: 2201.09760.
Wu, X., Chen, Q. G., Hu, Y., Wang, D., Chang, X., Wang, X., & Zhang, M. L. (2019). Multi-view multi-label learning with view-specific information extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 3884–3890).
Xiao, T., Reed, C. J., Wang, X., Keutzer, K., & Darrell, T. (2021). Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10539–10548).
Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., & He, L. (2022). Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16051–16060).
Xu, R., Huang, W., Zhao, J., Chen, M., & Nie, L. (2023). A spatial and adversarial representation learning approach for land use classification with POIs. *ACM Transactions on Intelligent Systems and Technology, 14*(6), 1–25.
Xu, Z., Chen, W., Zou, Y., Fang, Z., & Wang, S. (2024). Attention-based stackable graph convolutional network for multi-view learning. *Neural Networks, 180*, Article 106648.
Yan, W., Wu, M., Zhou, Y., Zheng, Q., Chen, J., Cheng, H., & Zhu, J. (2025). Label distribution-driven multi-view representation learning. *Information Fusion, 115*, Article 102727. Article.
Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*(4), 825–848.
Yao, Z., Fu, Y., Liu, B., Hu, W., & Xiong, H. (2018). Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 3894–3899).
Yong, X., & Zhou, X. (2024). MuseCL: Predicting urban socioeconomic indicators via multi-semantic contrastive learning. arXiv preprint arXiv:2407.09523.
Yu, H., Bian, H. X., Chong, Z. L., Liu, Z., & Shi, J. Y. (2024). Multi-view clustering with semantic fusion and contrastive learning. *Neurocomputing, 603*, Article 128264.
Yu, J., & Jia, A. L. (2024). AGCL: Adaptive Graph Contrastive Learning for graph representation learning. *Neurocomputing, 566*, Article 127019.
Zhang, L., Long, C., & Cong, G. (2022a). Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering, 35*(9), 9031–9036.
Zhang, M., Li, T., Li, Y., & Hui, P. (2020). Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 611–618).
Zhang, M., Zhu, Y., Liu, Q., Wu, S., & Wang, L. (2022b). Deep contrastive multiview network embedding. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 4692–4696).
Zhang, P., Niu, Z., Ma, R., & Zhang, F. (2025). Multi-view graph contrastive representation learning for bundle recommendation. *Information Processing & Management, 62*(1), Article 103956.
Zhang, Y., Fu, Y., Wang, P., Li, X., & Zheng, Y. (2019). Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1700–1708).
Zhang, Y., Xu, Y., Cui, L., & Yan, Z. (2023). Multi-view graph contrastive learning for urban region representation. In *Proceedings of the 2023 International Joint Conference on Neural Networks* (pp. 1–8).
Zhang, Y., Huang, W., Yao, Y., Gao, S., Cui, L., & Yan, Z. (2024). Urban region representation learning with human trajectories: A multi-view approach incorporating transition, spatial, and temporal perspectives. *GIScience & Remote Sensing, 61*(1), Article 2387392.
Zhao, J., Chen, C., Zhu, Y., Deng, M., & Liang, Y. (2025). UniTR: A unified framework for joint representation learning of trajectories and road networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(12), 13348–13356.
Zhou, S., He, D., Chen, L., Shang, S., & Han, P. (2023). Heterogeneous region embedding with prompt learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(4), 4981–4989.
Zou, X., Yan, Y., Hao, X., Hu, Y., Wen, H., Liu, E., … Liang, Y. (2025). Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion, 113*, Article 102606. Article.